Analysis of a Microbial Community Metagenome from Uranium-Contaminated Groundwater

Christopher L. Hemme[1,2], Terry J. Gentry[1,3], Liyou Wu[1,2], Matthew W. Fields[4], Chris Detter[5,6], Kerrie Barry[5], David Watson[1], Christopher W. Schadt[1], Paul Richardson[5], Terry Hazen[7], James Tiedje[8], Eddy Rubin[5] and Jizhong Zhou[1,2]

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN[1], Institute for Environmental Genomics, University of Oklahoma, Norman, OK[2], Department of Soil & Crop Sciences , Texas A&M University, College Station, TX[3], Department of Microbiology, Miami University, Oxford, OH[4], DOE Joint Genome Institute, Walnut Creek, CA[5], Los Alamos National Laboratory, Los Alamos, NM[6], Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA[7], Center for Microbial Ecology, Michigan State University, East Lansing, MI[8]

N-030

DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT

OFFICE OF SCIENCE U.S. DEPARTMENT OF ENERGY

ESPP Environmental Stress Pathway Project

IEG Institute for ENVIRONMENTAL GENOMICS

MICHIGAN STATE UNIVERSITY

VIMSS Virtual Institute for Microbial Stress and Survival

MIAMI UNIVERSITY

MISSOURI

UNIVERSITY OF WASHINGTON

DIVERSA

Sandia National Laboratories

OAK RIDGE NATIONAL LABORATORY

Berkeley University of California

http://vimss.lbl.gov/

INTRODUCTION

Due to the uncultivated status of the majority of microorganisms in nature, little is known about their genetic properties, biochemical functions, and metabolic characteristics. Although sequence determination of the microbial community 'genome' is now possible with high throughput sequencing technology, the complexity and magnitude of most microbial communities make meaningful data acquisition and interpretation difficult. Therefore, we are sequencing groundwater microbial communities with manageable diversity and complexity (~10-400 phylotypes) at the U.S Department of Energy's Environmental Remediation Sciences Program (ERSP) Field Research Center (FRC), Oak Ridge, TN. The microbial community has been sequenced from a groundwater sample contaminated with very high levels of nitrate, uranium and other heavy metals and pH ~3.7. Sequence analysis of this groundwater sample based on a 16S rDNA library revealed 10 operational taxonomic units (OTUs) at the 99.6% cutoff with >90% of the OTUs represented by an unidentified γ-proteobacterial species similar to *Frateuria*. Additional OTUs were related to a β-proteobacterial species of the genus *Azoarcus*. Three clone libraries with different DNA fragment sizes (3, 8 and 40 kb) were constructed, and 50-60 Mb raw sequences were obtained using a shotgun sequencing approach. The raw sequences were assembled into 2770 contigs totaling ~6 Mb which were further assembled into 224 scaffolds (1.8 kb-2.4 Mb). Preliminary binning of the scaffolds suggest 4 primary groupings (2 *Frateuria*-like γ-proteobacteria, 1 *Burkholderia*-like β-proteobacteria and 1 *Herbaspirillum*-like β-proteobacteria). Genes identified from the sequences were consistent with the geochemistry of the site, including multiple nitrate reductase and metal resistance genes. A low level of strain diversity was observed in the sample, with little significant polymorphism detected in the ORFs studied. We hypothesize that the major adaptive response within the community following site contamination resulted from lateral gene transfer events within the community followed by adaptive evolution of individual genetic elements. These adaptive events likely triggered multiple selective sweeps within the populations that have reduced the strain heterogeneity of the community.

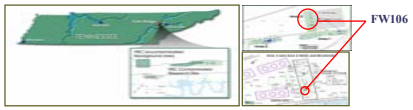


Fig. 1. Map showing the location of Area 3, well FW106 of the ERSP FRC site from which the microbial community genome sequences were obtained.

FW106 Groundwater Geochemistry

- Uranium – 51 mg/L (soil ~500 mg/kg)
- Nitrate – 2,331 mg/L
- Sulfate – 1997 mg/L
- Total Organic Carbon (TOC) – 244 mg/L
- Total Inorganic Carbon (TIC) – 284 mg/L
- pH – 3.7
- Heavy Metals and Organics – µg to mg/L

Metagenome Statistics

- ~70 Mb raw sequence
- ~8 Mb assembled sequence
  - 2770 contigs
  - 224 scaffolds (1.8 kb-2.4 Mb)
- 5 preliminary phylotypes based on identification of anchor genes (16S rRNA, 23S rRNA, *gyrB*, *recA*, *rpoB*, *ileS*, *fusA*)
- ~70 of metagenomic sequences assigned to these 5 phylotypes
- Estimated Genome Coverage*
  - FRC Gamma Group I (*Frateuria* I) – 9.7X
  - FRC Beta Group I (*Burkholderia*) – 1.1X
  - FRC Gamma Group II (*Frateuria* II) - <0.1X
  - FRC Beta Group II (*Herbispirillum*) - <0.1X
  - FRC Alpha Group I (*Afipia*) - <0.1X

* Genome coverage estimated using reference genome sizes as follows: *Burkholderia*, 7.0 Mb; *Herbispirillum*, 6.8 Mb; *Xanthomonas*, 4.2 Mb; *Bradyrhizobium*, 9.1 Mb

Significant Scaffolds

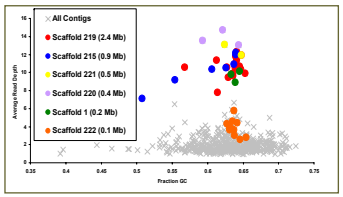


Fig. 2. Plot of Fractional GC vs. Average Read Depth for the 6 largest scaffolds from the FW106 metagenome. All of these represent species of the dominant γ-proteobacterial species (FRC γ Group I). The low read depth of scaffold 222 (orange) suggests that it may represent a separate γ-proteobacterial phylotype.

Phylogeny of FW106 Community

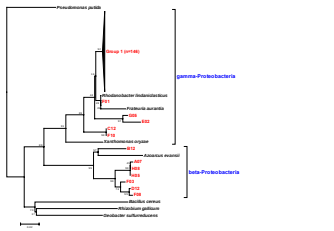


Fig. 3. Analysis of 16S rDNA from the metagenome sequence revealed 10 operational taxonomic units (OTU's) at a 99.6% similarity cutoff (2 mismatches). Phylogeny of the 16S sequences suggests that the community is dominated by a single γ-proteobacterial species similar to *Frateuria*. β-proteobacterial species were observed at lower proportions.

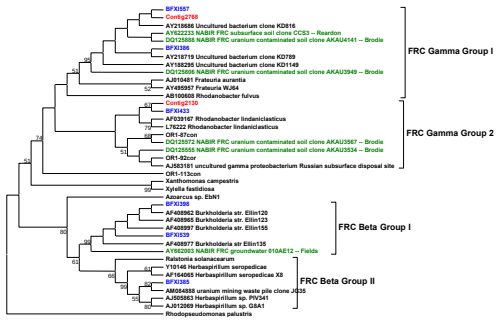


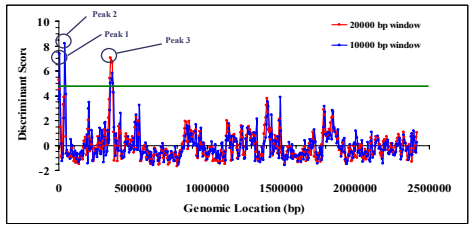
Fig. 4. Phylogeny of the predicted 16S rRNA sequences obtained from the assembled metagenomic sequence. The phylogeny defines the four primary phylotypes of the community and was used for preliminary binning of the remaining the contigs. Taxon labels are colored as follows: Red, genes predicted from the metagenome (2 gene fragments mapping to FRC Gamma Group I and FRC Beta Group I are not shown); Blue and Green, genes experimentally isolated from locations within the FRC.

Sampling of Genes from the Metagenome

- Energy Metabolism
  - 30+ Cytochromes (23 *c*-type)
  - 3 NiFe Hydrogenase complexes
  - Assimilatory and Dissimilatory Nitrate Reductase
  - Formate Dehydrogenase
  - Multiple Cytochrome Oxidase Complexes
  - Hexose and Pentose Transporters, including multiple PTS Transport Complexes
- Stress Response
  - RecA
  - CRP/FNR
  - Fur
  - Sigma Factors (including σ^{32} and σ^{38})
  - 30+ Sensor Histidine Kinases
  - 40+ cation efflux/multidrug resistance pumps
  - 7 Na^+/H^+ Antiporters
  - 9 Heat Shock Response Genes
- Metal Resistance Genes
  - Arsenate Reductase Operon
  - Mercuric Reductase Operon
  - Copper Resistance Genes
  - CoZnCd Transport Genes
  - Chromate Transport Genes
  - Misc. Heavy Metal Resistance Genes

Genes identified from the metagenome are consistent with those predicted based on the geochemistry of the site, i.e. metal resistance genes, denitrification genes, stress response systems.

Evidence for Lateral Gene Transfer



- Peak 1
  - ABC-type multidrug transport operon
  - Histidine kinase/CheY-like response regulator pair
  - Transposase
  - Hypotheticals
- Peak 2
  - RecF
  - CheY-like receiver
  - Hypotheticals
- Peak 3
  - Multiple alcohol dehydrogenase genes
  - Amino acid transport/metabolism genes
  - Transposase
  - Hypotheticals

Fig. 5. Identification of putative genomic islands (GI) in the major scaffold 219 (FRC γ Group I, 2.4 Mb). GI's were determined using iterative discriminant analysis as implemented in Tu and Ding, 2003 based on G+C difference, genome signature contrasts and codon usage bias. Analyses were conducted using sliding windows of 10000 (blue) and 20000 (red) bp. Peaks exceeding a discriminant score > 3.9 were considered to be potential GI's. CDS sequences in the regions of the predicted GI's are listed below the figure.

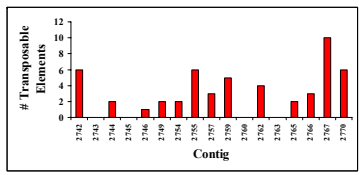


Fig. 6. Distribution of transposase, integrase, insertion elements and phage recombinase genes within the contigs of the primary scaffold 219. 52 total transposable elements were identified for this scaffold.

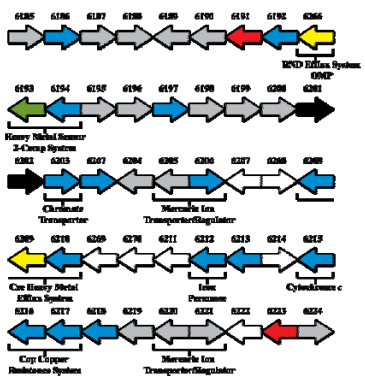


Fig. 6. The genomic environment of contig 2766 of scaffold 219 suggests evidence of lateral gene transfer of metal resistance genes into the γ groups from the β groups. ORFs are colored as follows: grey, γ-proteobacteria; blue, β-proteobacteria; yellow, α-proteobacteria; green, δ-proteobacteria; black, non-proteobacteria; white, unknown; red, transposase/integrase. ORFs are not to scale. Assignments were based on best Blast hit from the automatic annotation.

Evolution of the FW106 Community

		Fixed	Polymorphic
RecA	Nonsynonymous	11	1
	Synonymous	49	0
GroEL	Nonsynonymous	4	1
	Synonymous	29	0
Heavy Metal Sensor Histidine Kinase	Nonsynonymous	14	1
	Synonymous	30	1
Heavy Metal Efflux OMP, CzcC	Nonsynonymous	56	6
	Synonymous	94	3

Fig. 5. A sampling of the degree of polymorphism of genes located on the primary scaffold 219. ORF nucleotide sequences were compared to the metagenome read library using Blastn. Returned reads were clustered based on a 94% sequence identity cutoff and the resulting clusters were used to identify polymorphic sites. Sites were only considered to be polymorphic if at least two reads showed the same nucleotide change.

Based on visual analysis of the contig sequences using Consed and detailed analysis of selected individual genes, it appears that the level of detectable polymorphism in the sample is very low. The abnormal level of polymorphism in the examined czcC gene suggests a possible target of positive selection. The accumulated evidence suggests that the community has undergone multiple selective sweeps over the past 50 years that have significantly reduced the strain diversity of the community. The major adaptive events responsible for this phenomenon are likely the result of beneficial lateral transfer of genes between community members followed by/in conjunction with positive selection at the genetic level.

Summary

- Groundwater community is dominated by *Frateuria*-like γ-proteobacteria
- Preliminary analysis suggests low strain diversity within the established phylotypes, possibly as the result of selective sweeps within the population
- Denitrifying β-proteobacteria may serve as keystone species
- Metabolic and regulatory genes predicted to exist based on geochemistry were identified, suggesting that the community has undergone adaptive evolution in response to contamination of the environment
- Lateral gene transfer between community members followed by positive genetic selection is proposed as the major mechanism by which the community has adapted to the environmental pressures imposed on it

Future Work

- Metagenome-wide analysis of polymorphism
- Metagenome-wide survey of positive selection
- Metagenomic analysis of microbial communities from additional wells, including from the uncontaminated background site (proposed)
- Genomic sequencing of isolates of the major groups to test hypotheses on lateral gene transfer, positive selection, etc.

ACKNOWLEDGEMENT

ESPP is part of the Virtual Institute for Microbial Stress and Survival supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.